Z 学校编码：10384　　分类号_____密级____
学号：200326097　　UDC _____

厦　门　大　学

理　学　硕　士　学　位　论　文

疾病相关的选择性剪接基因与密码子和互补密码子的生物信息学分析

Bioinformatic Analysis of Disease-Associated Alternative Splicing genes and Codon-Complementary Codons

黄惠芳

指导教师姓名：纪志梁　副教授
马　飞　教　授

专　业　名　称：生化与分子生物学

论文提交日期：2006 年 4 月 28 日
论文答辩时间：2006 年 5 月 30 日
学位授予日期：2006 年　月　日

答辩委员会主席：陶涛 教授
评　阅　人：_____

2006 年 5 月

目　录

摘 要

遗传密码子(Genetic Codes)是连接基因和蛋白质的桥梁。通过研究密码子与互补密码子在疾病选择性剪接形成及物种进化中使用频率的差异，将有助于从新的角度来了解选择性剪接的进化及其导致人类疾病的可能分子机制。本文统计分析了 118 个人类疾病相关的选择性剪接基因编码区中 2,903 个单核苷酸替代的特征：在选择性剪接基因的单碱基替代中，最普通的替代类型依次是 C|T (C/T 或 T/C) (31.97 %) 和 A|G (28.14 %)替换、G|T (13.65 %) 和 C|G (12.47 %)， A|C 和 A|T 替换是较少的，分别为 7.44 %和 6.34 %。转换（A|G 和 C|T）60.11 %大于颠换（A|C、G|T、A|T 和 C|G）39.89 %。密码子第一、第二和第三位碱基替代的等次分别为 C > G > A > T、 G > C > T > A 和 C > G >T > A，其比例分别为 49.26 %、42.16 %和 8.58 %。在对 60 种生物 (13 种真核生物，26 种细菌，5 种古细菌和 16 种病毒) 基因组和 60 种细胞器 (40 种线粒体和 20 种叶绿体) 基因组的密码子与互补密码子数据的分析中，显示密码子与互补密码子存在显著或极显著的相关性。对 70 个物种（5 个古细菌，20 个真核生物和 45 个细菌）基因组密码子与互补密码子的分析支持了上一步的结果，并发现密码子与互补密码子的使用相关系数和 GC 含量之间存在显著的相关性。基于密码子与互补密码子使用频率差异，我们构建了密码子与互补密码子使用的进化树；结果表明密码子与互补密码子使用的进化树可以反应物种的协同进化过程。

关键词：选择性剪接；疾病；密码子使用偏爱

Abstract

Genetic code is the biological language in communication between DNA and protein. To reveal the role of different usage of genetic codons and their complementary codons in the evolution of disease-associated alternative splicing, 118 alternatively spliced human genes with 2,903 nucleotide substitutions were statistically analyzed. It is found that nucleotide tranitions has overall preponderance over nucleotide transversions (60.11 % versus 39.89 %). The propobility of substituation in the first, second and third base of codons are 49.26 %, 42.16 % and 8.58 %, respectively. Among tri-nucleotides (triplets) formed by the substitutions (denoted as N) and their immediate neighboring nucleotides (5' to 3'), 1,866 (64.28 %) are of pyrimidine-N-purine or purine-N-purine types. Futher exon analysis of these disease-related genes found that single base substitutions are much denser within the first and last two exons than that of middle exons. 83.91 % of substitutions have occurred within exons that have a medium range (40 ~ 65 %) of GC content. Larger exons seem to have suffered fewer substitutions. Additional study of 44 cellular organisms, 16 viruses, 40 mitochondria and 20 chloroplasts indicates that there exist significant correlation between codons and corresponding complementary codons usage in species except virus. Moreover, phylogentic tree was generated based on this finding, which is comparable to traditional molecular evolution tree.

Keywords: codon and complementary codon; Alternative splicing; Single base substitution

第一章 文献回顾

1 选择性剪接的研究进展

1.1 选择性剪接的发现

根据遗传的中心法则，遗传信息从DNA传递至mRNA，再由mRNA翻译成蛋白质。在原核生物中，基因的DNA序列全部转录成mRNA，mRNA序列又完整地翻译成蛋白质的氨基酸序列。相对而言，真核生物的遗传物质传递却复杂得多。1977年，Phillip Sharp和Richard Roberts[1,2]都发现了断裂基因，即真核细胞腺病毒mRNA序列并非来源于完整的DNA序列，该病毒的基因是断裂的；同时，Sharp等[1]发现mRNA上有不同的片段得以翻译，并根据其顺序性推测这些RNA片段存在某种新的连接方式，后来被证实为前体mRNA的剪接过程。Sharp和Roberts的发现顿时引起了RNA研究的热潮，在那之后，Walter Gilbert[3]提出了外显子和内含子的概念：基因中被编码的序列为外显子，不能被编码的序列为内含子。并指出同一基因的不同的外显子在被剪切之后，由于重组的方式不一样，会形成不同的mRNA（选择性剪接）。迄今为止，已经发现的mRNA前体剪接方式有两种：组成性剪接(constitutive splicing ,CS) 和选择性剪接(alternative splicing ,AS) 。组成性剪接的外显子存在于同一基因产生的所有mRNA中，而选择性剪接的外显子仅存在于该基因的部分mRNA中[4]。

在研究RNA的热潮中，有关选择性剪接的研究越来越广泛，从性别决定到细胞凋亡均有涉及[5]，几乎囊括了整个生命的历程。尤其是研究中发现了许多选择性剪接错误导致的疾病，从遗传性疾病到肿瘤，甚至是精神分裂症等获得性疾病，使选择性剪接在分子生物学和生物信息学等领域都倍受瞩目。

1.2 选择性剪接的模式与机制

从选择性剪接的发现开始，就不断地有人致力于研究其作用机制，然而，在不同的真核生物基因中，外显子和内含子的长度往往有很大差别，增加了选择性剪接机制研究的难度，导致许多选择性剪接的详细机制仍不清楚。在人类基因组中，外显子的平均长度为150个核苷酸，而内含子的平均长度为3500个核苷酸，最长的达到50万个核苷酸[6]，外显子之间通常都被很长的内含子所间隔。对于大内含子基因，剪接因子寻找外显子两侧正确的5’和3’剪

接位点，并形成剪接复合体，这一过程称为外显子限定(exon definition) [7] 。

典型的多外显子前体mRNA中，不同的剪接模式可以产生许多不同的mRNA。选择性剪接时由于剪接位点的改变，有的外显子被延长，有的被缩短，有的表达，有的不表达，而内含子可能被剪切或插入到mRNA中[6, 8]。如图1-1所示：图A表示的是最常见的选择性剪接模式，中间的外显子可能被表达，也有可能被跳过（Exon Skipping）；图B显示在多个外显子同时存在，且仅有一个能保留到mRNA中时，其他的被剪切掉；C图和D图显示对外显子上5’或3’端的剪接位点进行选择性剪接，使外显子延长或缩短；E图和F图分别显示剪接促进因子和Poly(A)对5’或3’ 端进行选择，从而改变mRNA的起始位点和终止位点；G图显示的是一个被保留的内含子，仍然存在被切除的可能；在H图中，前体mRNA中存在多个剪接位点，在剪切之后，不同的外显子可能按不同的结合方式产生许多不同的成熟mRNA，并且进一步翻译形成同一基因的不同表达产物，即蛋白质家族。

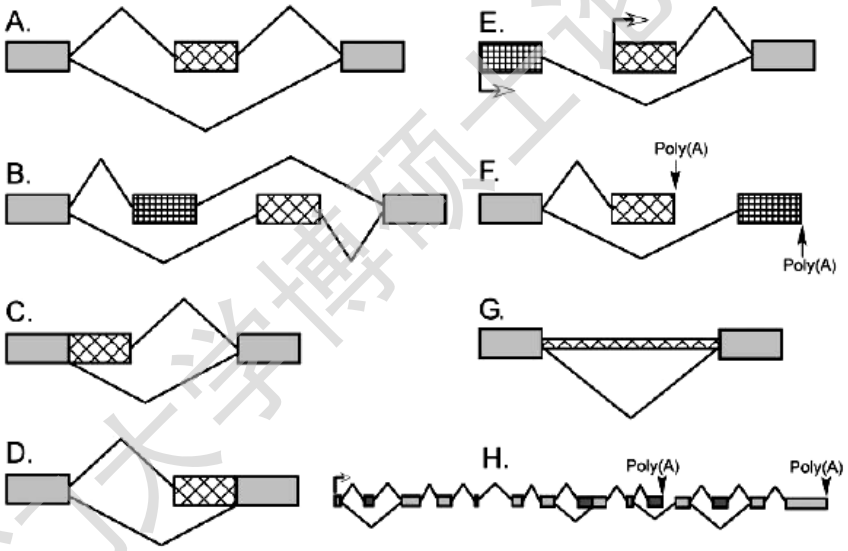

图1-1 选择性剪接的模式.（选自文献[8]）灰色表示组成性剪接的外显子，从始至终都存在于mRNA中，有网格线的表示选择性剪接的外显子，它们可能存在于mRNA中，也可能被剪切掉。

Fig 1-1 Patterns of alternatvie splicing. Constitutive sequences present in all final mRNAs are gray boxes. Alternative RNA segments that may or may not be included in the mRNA are gridding boxes.

由于mRNA加工时存在多种选择的可能性，选择性剪接的具体过程十分

复杂，虽然人们对选择性剪接的详细过程的了解仍然十分有限，但多数选择性剪接过程还是存在大致相似的机制。一般来说，前体mRNA都有着相似的结构。选择性剪接的正确进行很大程度地依赖于剪接位点的正确识别，所谓剪接位点，即外显子和内含子相连接的地方，其内含子的5’端以一个GU双核苷酸为起点，引导一段较长但保守性较差的序列。选择性剪接位点的改变可能会使蛋白质产物发生各种各样的后果：在氨基酸序列上的小小改变可能会导致配基改变、酶活性变化、变构现象和蛋白质定位的变化等，而这些事件的发生在生物体上的表现即为疾病。内含子的3’端包含三个保守元件：分支点、多聚嘧啶区和一个AG双核苷酸的结尾，其中AG是内含子3’末端的标志[8]。外显子上有许多保守区域，如外显子增强子（exonic splicing enhancers，ESEs），能促进剪接因子结合到前体mRNA上；与之作用相反的是外显子沉默子（exonic splicing silencers, ESSs）。相应地，内含子中也存在内含子增强子（Intronic splicing enhancers，ISEs）和内含子沉默子（Intronic splicing silencers, ISSs）[6]。在真核生物前体mRNA剪接过程中，非剪接位点的调节序列严重地干扰选择性剪接的正常进行，剪接增强子和沉默子对选择性剪接位点的正确识别起了很重要的作用。

选择性剪接的执行者被称为剪接机器或剪接体（Spliceosome）[9]。剪接体是由5种核内核糖体蛋白（U1，U2，U4，U5，U6snRNP）和超过200种的剪接因子组成的大分子复合物，它能结合到内含子上，催化剪接中的两步关健的转酯反应。第一步反应时，分支位点处的腺苷酸2’羟基亲核攻击5’剪接位点的3’-5’磷酸二酯键，剪接出两种中间产物：一个是游离的5’端的外显子；另一个是内含子与3’端外显子形成的套索状分子，它们之间通过2’-5’磷酸二酯键连接。第二步反应是由游离5’端的外显子发起的：外显子的3’羟基攻击内含子3’剪接位点上磷酸二酯键，去除套索结构上的内含子序列，同时将相邻的两个外显子连接起来[10]。

在组成剪接体的200多种蛋白中，包含了富含精氨酸和丝氨酸的SR蛋白，hnRNP、RNA螺旋酶、激酶等许多不同的蛋白质。剪接复合体刚开始装配时，U1snRNP结合至5’剪接位点，辅助因子U2AF结合至3’剪接位点。U2AF是U2snRNP的辅助因子，包含了两个亚基，其大小分别为65kDa和35 kDa，

U2AF65识别多聚嘧啶区域，U2AF35识别3’剪接位点的AG。在U2AF结合之后，分支点结合蛋白SF1（splicing factor 1，在酵母中为BBP）结合到内含子的分支点上，形成了剪接体的初始形态－E复合体[10, 21]，如图1-2。随后U2snRNP也结合到分支点上，从而形成了剪接体的另一中间状态－A复合体。A复合体结合U4/U5/U6 snRNP之后形成B复合体， B复合体经过复杂的重排之后形成C复合体，即成熟的剪接体。在成熟的剪接体中，5’剪接位点的U1snRNP被U6snRNP所替代，而U1和U4则脱离了剪接体[8]，这时，剪接体便可催化剪接中的两步关健的化学反应。

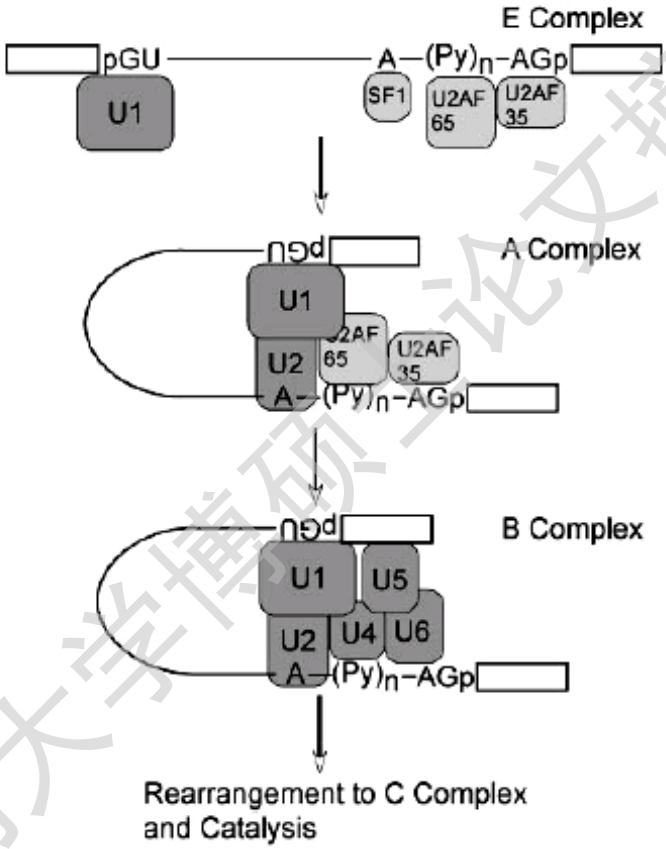

图1-2　剪接体的装配过程.（引自文献[11]）

Fig 1-2 The spliceosome assemble onto the intron.

剪接体装配过程的变化会引起剪接位点的变化。在多数情况下，剪接位点的选择是由起始因子（与前体mRNA结合）和早期的剪接体复合物所调节的。这时E复合体已形成，并以碱基配对原则结合到前体mRNA的剪接位点上，相应的内含子已在剪接范围内。然而，这种剪接位点的识别方式并不是绝对的，在Lallena[12]等的研究中发现，3’端的剪接位点选择是在第一次转酯反应

之后进行的。与此同时，剪接位点的序列特征并不能完全决定这一位点是否能与剪接体结合，并进行选择性剪接。前体mRNA上的外显子或内含子增强子（ESEs或ISEs）、外显子或内含子沉默子（ESSs或ISSs）对选择性剪接的位点的正确识别能起一定的作用，外显子限定(exon definition) 也能促进剪接位点的正确识别，除此之外，调节蛋白对剪接位点的识别也起了重要的作用[13, 14]。

1.3 选择性剪接与人类疾病

目前的研究发现，许多疾病都是由于前体mRNA剪接错误导致的，剪接位点的改变、错误的调控过程都会导致疾病的产生。根据人类基因突变数据库（Human Gene Mutation Database）统计的结果，在人类突变的基因中，有10%是由于选择性剪接位点突变引起的[15]。按照外显子的平均长度为 300 个核苷酸计算，在前体mRNA的 3’剪接位点附近，大约每隔 20 个碱基就有一个秘密的剪接位点（Critical site site），5’端附近大约每隔 10 个碱基就有一个秘密剪接位点。然而，这一推测并不包括ESE/ISE和ESS/ISS在内，剪接引起的突变仍然很可能被低估了[16]。

在不同类型的选择性剪接疾病当中，突变发生的情况不一样，导致的选择性剪接错误也不尽相同。首先，突变会导致一些意外的剪接功能的获得，如产生秘密剪接位点，ESE/ISE和ESS/ISS等[16]。最早发现突变产生 3’端秘密剪接位点是在β球蛋白基因中，该突变会导致地中海型贫血症（Beta+-thalassemia），表现为严重的贫血甚至死亡[17]。脊柱肌肉细胞萎缩（Spinal muscular atrophy，SMA）是一种常染色体隐性遗传病，其发病原因是纯合子缺乏SMN1 基因，为选择性剪接时第七外显子被跳过所致，然而，SMN1 的同源基因SMN2 所产生的SMN蛋白会增加发病的可能性。SMN2 和SMN1 非常接近，但SMN2 的第七个外显子第六位的C转换成T，并因此产生一个ESS，从而增加了该外显子被跳过的机率[18]。与SMN基因类似，编码丙酮酸脱氢酶的基因突变并使之获得ESE，导致外显子限定发生错误，产生功能不完善的丙酮酸脱氢酶，因此导致发育延迟和乳酸酸中毒(lactic acidosis)[19]。

其次，突变也会导致剪接功能的缺失，具体表现为顺式元件的失活。因

此，许多遗传病的产生都是正常的剪接位点发生突变的结果。当正常的剪接位点被中断时，ISE和ESE都可能引起外显子跳过或外显子的改变，从而导致疾病的发生[20]。例如，B型血友病是由于缺乏第九生长因子导致的，其原因是第九生长因子的外显子f下游+1 位置有一个G发生颠换，突变成T，导致血友病的发生[21]。此外，还有一典型例子为家族性、孤立性生物激素缺乏Ⅱ型遗传病（familial isolated growth hormone deficiency type II, IGHD II）。IGHD II是由于GH-1（垂体前叶生长激素）基因突变导致的显性遗传病，患者主要特征为身材异常矮小。GH-1 基因有 5 个外显子，同时还存在一定比例的选择性剪接mRNA（5%-10%）。GH-1 的全长蛋白质分子量为 22 kDa，然而，3’剪接位点的选择性剪接使其产生另外两个蛋白质：一个是分子量为 20kDa的蛋白质，为三外显子被剪切了 45 个核苷酸所致；另一种蛋白分子量为 17.5 kDa，在选择性剪接时第三外显子整个被跳过。所有的IGHD II突变均由于剪接元件被中断，如顺式作用元件ISE或ESE，或 5’剪接位点的中断。任意一个剪接元件被中断，都会使 3’剪接位点的选择性剪接增加。在本例中，当第三外显子的第五位核苷酸由G突变成A时，打断了一个正常的ESE，从而使整个第三外显子被跳过，产生 17.5 kDa的蛋白质，导致疾病的发生[22]。

除了突变导致的剪接功能获得与缺失外，突变可能会发生在组成剪接体的剪接因子中，或剪接因子的调节因子和辅助因子中。在酵母中，无义突变发生剪接体的组成元件中通常都是致死的，而在多细胞动物中，至少在细胞水平上是致死的[20]。目前已发现许多剪接因子突变导致的人类疾病，视网膜色素变性（Retinitis pigmentosa，RP）便是这一类型疾病的典型。视网膜色素变性是一种具有遗传异质性的疾病，其发病率约为 1/4000，主要特征为视网膜色素进行性变性，夜盲、周边视野缺失以致最后全盲。它可以表现为常染色体显性，也可表现为常染色体隐性遗传或X连锁遗传。目前已有 30 个不同的RP基因及连锁位点被确定，其中PRPF31、HPRP3 和PRPC8 基因是引起常染色体显性遗传性视网膜色素变性的致病基因。在这三个引起显性遗传性RP的剪接因子基因中，PRPF31 是目前了解得最清楚的一个。人类PRPF31 是一种U4/U6snRNP与U5snRNP的关联蛋白，它能与一种大小为 102kDa的U5 特异性蛋白直接作用，促进U4/U6snRNP与U5snRNP的联合[23]，从而使剪接体

的装配得以进行。PRPF31 发生突变时导致这一步骤无法进行，从而导致常染色体显性遗传性视网膜色素变性，其突变的类型包括插入、缺失、无义突变和剪接位点的突变[24]。

RP是由于剪接因子突变，并导致剪接体装配中snRNP的结合被中断产生的。而强直性肌营养不良（Myotonic dystrophy，DM）则是由于选择性剪接调节因子的功能丧失引起的。其主要症状为骨骼肌张力增高、进行性肌坏死、心血管畸形等。目前已发现的有两种DM，最常见的DM1 是由位于 19q13.3 上DM蛋白激酶基因（DMPK）的 3'UTR的CTG重复数增加所致，正常个体的重复次数<40 次，DM患者在 80 次至数千次之间。分析发现这类病人的RNA中含有从扩展的DMPK基因转录而来的长的CUG重复，后者影响了CUG重复结合蛋白（CUG-repeat-binding protein, CUG-BP）的功能，而CUG-BP是选择性剪接中的调节因子[25]。 在DM1 受损的肌肉组织中CUG－BP的水平增高，结果使其作用的靶基因心肌肌钙蛋白T（cardiac troponin T, CTNT）基因、肌肉特异性氯离子通道 1 基因（muscle-specific choride channel 1, CIC-1）、胰岛素受体基因（insulin receptor, IR）3 个基因的mRNA前体出现异常的选择性剪接：CTNT保留第五外显子，IR的第十一外显子跳过，而CIC则保留了第二内含子[26]。

除了上述几种导致选择性剪接疾病的原因外，目前尚有许多疾病的详细机制仍不清楚。虽然选择性剪接可能对某些获得性疾病有着至关重要的作用，如肿瘤、缺血症等，但由于疾病本身的症状不稳定，或与选择性剪接的联系不够确定，至今仍无法做出明确的分类，相关疾病的机制仍在研究中。总的来说，选择性剪接出错会导致各种各样的疾病，从某一器官肿瘤到多个系统参与的综合症都有，剪接突变导致的疾病在人体各大系统中都有分布。如表 1-1 所示[15]。

表 1-1　选择性剪接导致的人类疾病

Table 1-1 Human diseases associated with aberrant or defective alternative splicing.

类型	疾病	基因	剪接错误
肿瘤	肺癌 (Lung cancer)	p53, p51, CD44, EGFR	剪接产物异常或缺乏
	不同组织的癌症 (Cancer in different tissues)	p53, CD44	剪接产物异常或缺乏
	Wilm's 瘤 (Wilm's tumor)	CD44	Aberrant splicing isoforms
	乳腺癌 (Breast cancer)	HER2/neu	剪接错误及剪接中间物不平衡
	乳腺和卵巢癌 (Breast and ovarian cancer)	BRCA1, BRCA2	外显子跳过
	乳腺癌、横纹肌瘤 (Breast cancer, rhabdomyosarcoma)	Mdm2	致瘤的变体缺乏 P53 结合域
	纤维神经瘤 (Neurofibromatosis type 1)	NF-1	外显子跳过
	成神经细胞瘤 (Neuroblastoma)	HUD, HUC, NP-1, α-internexin	产生肿瘤抗原
心血管疾病	血胆脂醇过多 (Hypercholesterolemia)	LDLR	使用秘密的剪接位点，外显子跳过
	高甘油三脂血症 （Hypertriglyceridemia）	肝脂肪酶	使用秘密的剪接位点
	MARFAN氏综合症 （Marfan syndrome）	Fibrillin-1	使用秘密的剪接位点，外显子跳过
	原发性心肌症 (Cardiomyopathy)	cTNT	剪接的中间产物异常
	高血压 (Hypertension)	G-蛋白 β3	外显子跳过
代谢类疾病	II 型糖原贮积症 （Glycogen storage disease, type II）	Lysosomal α-glucosidase	使用秘密的剪接位点
	遗传性酪氨酸血症 (Hereditary tyrosinemia, Type I)	Fumarylaceto acetate hydrolase	外显子跳过
	间歇性卟啉症 (Acute intermittent porphyria)	胆色素原脱氨酶	外显子跳过
	血浆铜蓝蛋白不足 (Ceruloplasmin deficiency)	血浆铜蓝蛋白	使用秘密的剪接位点

	Fabry's 病 （Fabry's disease）	Lysosomal α galactosidase A	使用秘密的剪接位点
	Tay-Sachs 病 （Tay－Sach's Disease）	β-己糖胺酶	内含子保留，外显子跳过
	Sandhof 病 （Sandhof disease）	β-己糖胺酶β亚基	使用秘密的剪接位点
神经系统疾病	17 号染色体相关的额颞叶痴呆和帕金森综合征（FTDP-17）	Tau	剪接中间物不平衡
	Alzheimer's 病 (Alzheimer's disease)	Presenilin-1/ Presenilin-2	外显子跳过
	共济失调 (Ataxia telangiectasia)	ATM	使用秘密的剪接位点
	多发性硬化 (Multiple sclerosis)	CD45	剪接中间物不平衡
	脊柱肌肉细胞萎缩 (Spinal muscular atrophy)	SMN1, SMN2	使用秘密的剪接位点，外显子跳过
	视网膜色素变性 (Retinitis pigmentosa)	HPRP3, PRPF31,PRPC8	未知
精神异常	精神分裂症 (Schizophrenia)	$GABA_AR_\gamma$ 2,NCAM	剪接中间物不平衡
		NMDAR1	异常的选择性剪接
	注意力缺陷过动症（ADHD）	Nicotinic acetylcholine receptor	异常的选择性剪接
其他	囊肿性纤维化 （Cystic fibrosis）	CFTR	外显子跳过
	家族性，孤立性生长激素缺乏II型遗传病（IGHD II）	GH-1	剪接中间物不平衡
	Frasier 综合症 （Frasier syndrome）	WT 1	剪接中间物不平衡
	癫痫症 (Epilepsy)	AMPA 受体	剪接中间物不平衡
	门克斯病 （Menkes disease）	MNK	外显子跳过
	地中海型贫血 （Beta-thalassemia）	β球蛋白	使用秘密的剪接位点
	脑白质营养不良 (Metachromatic leukodystrophy)	Arylsulfatase A	使用秘密的剪接位点
	强直性肌营养不良 (Myotonic dystrophy)	DMPK	异常的剪接

注：资料来源：http://www.uwcm.ac.uk/uwcm/mg/, 2006 年 3 月。

1.4 选择性剪接与分子进化

正如前文所述，选择性剪接是广泛存在于真核生物中的一种前体 mRNA 加工机制，它能大大增加转录组的复杂性，使不同的转录过程在特定的时间里在不同的组织中表达，同时，选择性剪接还与许多疾病的发生有关，使之有望成为疾病诊断和治疗的有力工具之一。然而，选择性剪接从哪里来，它在自然选择中有着怎样的优势，又是怎么进化成这样一种机制？对于这一系列的问题，科学家们仍然无法做出明确的回答，但选择性剪接在进化方面的研究已经成为一个十分重要而有趣的方向。

在基因组的进化中，内含子的出现是一个非常重要的现象，因此，许多人试图通过内含子预测选择性剪接的早期形态，从而进一步了解选择性剪接的起源。众所周知，原核生物编码蛋白的基因中没有内含子，不存在选择性剪接；而真核生物虽然都有内含子，但选择性剪接只有在较高等的多细胞生物中才广泛存在。例如，酿酒酵母（*Saccharomyces cerevisiae*）的基因中只有大约 3%含有内含子，且只有 6 个基因中含有 2 个内含子；而酵母家族中的另一成员，裂殖酵母（*Schizosaccharomyces pombe*），约 43%的基因存在选择性剪接，并且许多基因中含有多个内含子，然而，该物种迄今仍未发现选择性剪接[27]。总的来说，在酵母中约存在 230 个内含子，但只有 3 个存在选择性剪接[7]，有人认为在酵母的许多基因中存在内含子丢失[28]，因此，从酵母中预测选择性剪接的早期形态被认为是不合适的。虽然如此，目前的预测中仍然很大程度地采用了酵母的RNA序列。对选择性剪接的早期形态进行预测，通常采用选择性剪接长度差异序列（length difference alternative splicing，LDAS）来与原核生物或酵母中的同源序列进行比对，当长的LDAS片段与原核生物或酵母中的同源序列相匹配，而短的LDAS片段出现明显的沟（gap）时，长的LDAS片段被认为是选择性剪接的早期形态；当短的LDAS片段与原核生物或酵母中的同源序列相匹配（没有沟），而长的LDAS片段出现沟时，则说明短的LDAS在进化时被插入了一些短的片段[29]。在多序列比对时，如果所有的序列匹配性都很高的话，所选的LDAS可能是来源于选择性剪接早期形态的某一特定的蛋白质结构域。如图 1-4（a），在人类去氧亥普酸合成酵素（deoxyhypusine synthase）中，长的LDAS被认为是选择性剪接的早期状

态。从图中可以看出LDAS与原核生物和酵母中的同源序列的匹配程度非常高；图b显示的是大鼠和鸡的酪氨酸激酶的长LDAS片段的比对。其中鸡和大鼠匹配程度较高，为选择性剪接的早期状态，在酵母和细菌中出现明显的沟，说明该基因在进化时曾经被插入短的序列。

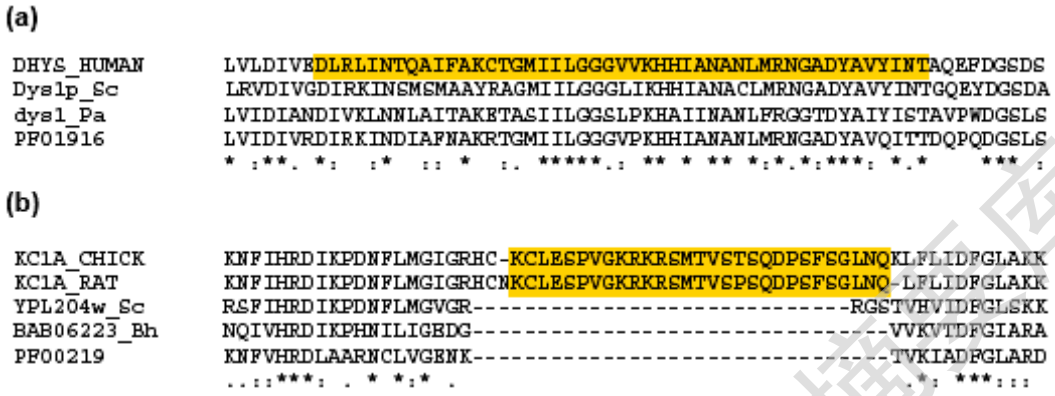

图 1-4 选择性剪接长度差异.（引自文献[29]）高亮部分显示的是选择性剪接的外显子。__后的文字代表不同物种：HUMAN 为人类，Sc 为酿酒酵母，Pa 为热球菌，Bh 为耐盐芽孢杆菌。星号表示不变的残基，冒号表示非常相似的残基，而点表示相似性一般的残基。

Fig 1-4 Length difference alternative splicing. The alternative exon sequences are highlighted. Asterisks show invariant residues, colons show positions occupied by very similar residues and full stops show somewhat similar residues in all aligned sequences. Species abbreviations: Sc, *Saccharomyces cerevisiae*, Pa, *Pyrococcus abyssi*, Bh, *Bacillus halodurans*.

除了对选择性剪接的早期形态进行预测外，人们试图通过选择性剪接的特点来了解其进化机制。目前关于选择性剪接的进化机制主要有两种观点，一个是以序列为基础，另一个则以选择性剪接调节因子为前提。前一种观点认为进化是DNA序列上的突变造成的：突变产生较弱的剪接位点，在许多剪接事件同时存在时，剪接体很可能跳过内部的外显子，使该细胞产生具有新功能的转录子。研究表明，选择性剪接的位点比组成性剪接弱，这种情况下外显子识别可能会按照选择性剪接的方式进行[30-32]。Modrek和Lee[33]认为选择性剪接在进化中起了重要的作用，是因为它能减轻基因进化时外显子所面临的选择压力。当基因的功能变发生改变时，其序列必须经过一系列突变，而变化中的序列通常要降低其适应性，在长时间的突变累积过程中，外显子